Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period

Clément de Chaisemartin, Xavier d'Haultfoeuille, Félix Pasquier, Gonzalo Vazquez-Bare

Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period

Clément de Chaisemartin* Xavier D'Haultfœuille† Félix Pasquier
Gonzalo Vazquez-Bare

Abstract

We propose new difference-in-difference (DID) estimators for treatments continuously distributed at every time period, as is often the case of trade tariffs, or temperatures. We start by assuming that the data only has two time periods. We also assume that from period one to two, the treatment of some units, the movers, changes, while the treatment of other units, the stayers, does not change. Then, our estimators compare the outcome evolution of movers and stayers with the same value of the treatment at period one. Our estimators only rely on parallel trends assumptions, unlike commonly used two-way fixed effects regressions that also rely on homogeneous treatment effect assumptions. With a continuous treatment, comparisons of movers and stayers with the same period-one treatment can either be achieved by non-parametric regression, or by propensity-score reweighting. We extend our results to applications with more than two time periods, no stayers, and where the treatment may have dynamic effects.

Keywords: differences-in-differences, continuous treatment, two-way fixed effects regressions, heterogeneous treatment effects, panel data, policy evaluation.

JEL Codes: C21, C23

1 Introduction

A popular method to estimate the effect of a treatment on an outcome is to compare over time units experiencing different evolutions of their exposure to treatment. In practice, this idea is implemented by estimating regressions that control for unit and time fixed effects. de Chaisemartin and D'Haultfœuille (2020*a*) find that 26 of the 100 most cited papers published by the AER

*University of California at Santa Barbara, clementdechaisemartin@ucsb.edu

†CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr. Xavier D'Haultfœuille conducted part of this research at the Paris School of Economics, and thanks them for their hospitality.

from 2015 to 2019 have used a two-way fixed effects (TWFE) regression to estimate the effect of a treatment on an outcome. de Chaisemartin and D'Haultfœuille (2020*b*), Goodman-Bacon (2021), and Borusyak et al. (2021) have shown that TWFE regressions are not robust to heterogeneous effects: under a parallel trends assumption, those regressions may estimate a weighted sum of treatment effects across periods and units, with some negative weights. Owing to the negative weights, the treatment coefficient in TWFE regressions could be, say, negative, even if the treatment effect is positive for every unit $\times$ period. Importantly, the result in de Chaisemartin and D'Haultfœuille (2020*b*) applies to binary, discrete, and continuous treatments.

Several alternative difference-in-difference (DID) estimators robust to heterogeneous effects have been recently proposed. Some of them apply to binary treatments that follow a staggered design, meaning that once they get treated, units cannot switch out of the treatment (see Sun and Abraham, 2020; Callaway and Sant'Anna, 2020; Borusyak et al., 2021). Others apply to binary or discrete treatments that may not follow a staggered design (see de Chaisemartin and D'Haultfœuille, 2020*b*,*a*). Finally, other estimators apply to continuous treatments that follow a staggered design, meaning that all units start with a treatment equal to 0, may then get treated at different dates with differing intensities, but once a unit gets treated its treatment intensity never changes (see de Chaisemartin and D'Haultfœuille, 2020*a*; Callaway et al., 2021). This last set of papers does not consider the case where the treatment is continuously distributed at every period. The goal of this paper is to complement the literature, by proposing heterogeneity-robust DID estimators of the effect of such treatments. This extension is practically important: TWFE regressions have often been used to estimate the effect of treatments continuously distributed at every time period, such as trade tariffs (see Fajgelbaum et al., 2020) or precipitations (see Deschênes and Greenstone, 2007).

We assume that we have a panel data set, whose units could be geographical locations such as states or counties. We start by considering the case where the panel has two time periods. We assume that from period one to two, the treatment of some units, hereafter referred to as the movers, changes. On the other hand, the treatment of other units, hereafter referred to as the stayers, does not change. Our target parameter is the average, across all movers, of the effect of moving their treatment from its period-one to its period-two value scaled by the difference between these two values. In other words, our target parameter is the average slope of mover's potential outcome function, from their period-one to their period-two treatment, hereafter referred to as the average of movers' potential outcome slope (AMPOS). The AMPOS can be interpreted as an average effect of increasing the treatment by one unit on the outcome. Our estimator of the AMPOS first compares the outcome evolution of each mover to the outcome evolution of stayers with the same value of the treatment in period one. Then, it divides those comparisons by the mover's treatment evolution, and finally averages the resulting ratios across all movers. Under parallel trends assumptions on all potential outcomes, the corresponding

estimand identifies the AMPOS. Graham and Powell (2012) have also proposed an estimator of the AMPOS. We compare their approach to ours below.

When some movers experience a small treatment change from period one to two, estimators of the AMPOS may be noisy, owing to the small denominator of those movers' slope estimator (see Graham and Powell, 2012). Accordingly, we also consider a second causal effect. This effect is a weigthed average, across all movers, of the slope of each mover's potential outcome function from its period-one to its period-two treatment, where movers receive a weight proportional to the absolute value of their treatment change from period one to two. Hereafter, we refer to this causal effect as the weighted average of movers' potential outcome slope (WAMPOS). We propose two estimators of that effect. Our first estimator compares the outcome evolution of each mover to the outcome evolution of stayers with the same value of the treatment in period one, averages such comparisons across all movers, and finally divides that average by movers' average treatment evolution from period one to two. Our second estimator starts by comparing the average outcome evolution of movers and stayers, without conditioning on their period-one treatment but reweigthing stayers by propensity score weights that ensure that their period-one treatment distribution is the same as among movers. Then, it divides that average by movers' average treatment evolution from period one to two. Under the same parallel trends assumptions as above, the estimands attached to our two estimators identify the WAMPOS.

Overall, our key idea is to compare movers and stayers with the same value of the treatment at period one. With a continuous treatment, this can either be achieved by a non-parametric regression, as implemented in our estimator of the AMPOS and in our first estimator of the WAMPOS, or by propensity score reweighting, as implemented in our second estimator of the WAMPOS.

The estimators proposed so far require that there be some stayers, whose treatment does not change from period one to two. This assumption is likely to be met when the treatment is say, trade tariffs: tariffs' reforms rarely apply to all products, so it is likely that tariffs of at least some products stay constant over time. On the other hand, this assumption is unlikely to be met when the treatment is say, precipitations: US counties never experience the exact same amount of precipitations over two consecutive years. We show that our identification results can be extended to the case where there are no stayers, provided there are quasi-stayers, meaning units whose treatment barely changes from period one to two. This assumption is likely to hold when the treatment is, say, precipitations: for every pair of consecutive years, there are probably some US counties whose precipitations almost do not change.

Finally, we also extend our results to applications with more than two time periods, and where the treatment may have dynamic effects. We also discuss how our estimators can be applied to discrete treatments taking a large number of values.

As mentioned above, our paper is related to the recent literature on heterogeneity-robust DID estimators. The two most closely related papers are de Chaisemartin and D'Haultfœuille (2020a) and Callaway et al. (2021), who also propose DID estimators of the effect of a continuous treatment. Here, we assume that the treatment is continuously distributed at every period, while they assume that the treatment is continuously distributed at certain periods only. For instance, in the two-periods case they assume that all units have a treatment equal to zero in period one, while we assume that the treatment is continuously distributed in period one as well. Accordingly, our papers do not overlap and complement each other.

Our paper builds upon several previous papers in the DID literature. First, it is closely related to the pioneering work of Graham and Powell (2012), who also propose a DID estimator of the AMPOS when the treatment is continuously distributed at every time period. Their estimator compares the outcome evolution of movers to that of quasi stayers and stayers, without conditioning on units' period-one treatment. The estimator of the AMPOS we propose is similar to theirs, except that it conditions on units' period-one treatment. Their estimator and our rely on similar parallel trends conditions. We impose parallel trends conditions on all potential outcomes, which imply that units experience parallel evolutions of their treatment effects. The condition stated in Equation (6) in Graham and Powell (2012) also implies that units experience parallel evolutions of their treatment effects. On the other hand, their estimator also relies on a linear treatment effect assumption, while our estimator does not rest on that assumption.

The idea to compare movers and stayers with the same baseline treatment also appears in de Chaisemartin and D'Haultfœuille (2018), de Chaisemartin and D'Haultfœuille (2020b), and de Chaisemartin and D'Haultfœuille (2020a), who had used that idea to form DID estimators of the effect of a binary or discrete treatment. With a non-continuous treatment, there will often be movers and stayers with the exact same baseline treatment. With a continuous treatment, the sample will not contain movers and stayers with the exact same baseline treatment, so this paper's contribution is to use non-parametric regression or propensity-score reweighting to compare movers and stayers "with the exact same baseline treatment".

Finally, D'Haultfoeuille et al. (2021) also consider a DID-like estimator of the effect of a continuous treatment, but their estimator relies on a common change assumption akin to that in Athey and Imbens (2006) rather than on a parallel trends assumption, and it requires repeated cross sections rather than panel data. Our first identification result is also related to Hoderlein and White (2012), who show how to identify the average marginal effect of a continuous treatment with panel data. The main difference is that they rule out any systematic effect of time on the outcome.

After some relabelling, the first estimand we propose to identify the WAMPOS is equivalent to the W^X_{DID} estimand proposed in Theorem S4 of the supplement of de Chaisemartin and D'Haultfœuille (2018), to identify the effect of a "fuzzy" binary treatment under a conditional

parallel trends assumption. In the definition of our estimand, the mover indicator and the period-one treatment respectively play the role of the treatment-group indicator and of the covariates in the W^X_{DID} estimand. Obviously, even though those estimands can be shown to be equal after some relabelling, they identify different treatment effects under different assumptions in the two papers. Still, this connection implies that our first estimator of the WAMPOS can readily be computed using, say, the `fuzzy_did` Stata command (see de Chaisemartin et al., 2019), which computes the $\widehat{W}^X_{DID}$ estimator of W^X_{DID} proposed by de Chaisemartin and D'Haultfœuille (2018). Similarly, after some relabelling the numerator of the second estimand we propose to identify the WAMPOS is equivalent to the estimand proposed in Equation (10) of Abadie (2005), to identify the effect of a "sharp" binary treatment under a conditional parallel trends assumption. In the definition of our estimand, the indicator for being a mover and the period-one treatment respectively play the role of the treatment and of the covariates in Abadie (2005). Again, this connection implies that the numerator of our second estimator of the WAMPOS can readily be computed using, say, the `absdid` Stata command (see Houngbedji, 2016). Similarly, when we allow for dynamic effects, the reduced-form estimands we propose in Equation (7) (resp. (8)) is equal to a weighted average of the estimands proposed in Equation (2.6) (resp. Equation (2.5)) of Callaway and Sant'Anna (2020) to identify the instantaneous and dynamic effects of a binary and staggered treatment. In the definition of our estimands, the indicator for being a first-time-mover and the period-one treatment respectively play the role of the treatment and of the covariates in Callaway and Sant'Anna (2020). Again, this connection implies that the estimators attached to our reduced-form estimands can be computed using, say, the `csdid` Stata command (see Rios-Avila et al., 2021).

2 Set-up and assumptions

A representative unit is drawn from an infinite super population, and observed at two time periods. This unit could be an individual or a firm, but it could also be a geographical unit, like a county or a region.[1] All expectations below are taken with respect to the distribution of variables in the super population. We are interested in the effect of a continuous and scalar treatment variable on that unit's outcome. Let D_1 and D_2 respectively denote the unit's treatments at periods 1 and 2, and $\mathcal{D}_1$ and $\mathcal{D}_2$ be their support. For any $d \in \mathcal{D}_1 \cup \mathcal{D}_2$, let $Y_1(d)$ and $Y_2(d)$ respectively denote the unit's potential outcomes at periods 1 and 2 with treatment d. Finally, let Y_1 and Y_2 denote her observed outcomes at periods 1 and 2. Let $M = 1\{D_2 \neq D_1\}$ be an indicator equal to 1 if the unit's treatment changes from period one to two, i.e. if the unit is a mover. Let also $M_i = 1\{D_2 > D_1\}$ be an indicator equal to 1 if the unit's treatment increases

[1] In that case, one may want to weight the estimation by counties' or regions' populations. Extending the estimators we propose to allow for such weighting would be straightforward.

from period one to two, and let $M_d = 1\{D_2 < D_1\}$ be an indicator equal to 1 if the unit's treatment decreases from period one to two.

We make the following assumptions.

Assumption 1 *(Parallel trends) For all $d \in \mathcal{D}_1 \cup \mathcal{D}_2$, almost surely, $E(Y_2(d) - Y_1(d)|D_1, D_2) = E(Y_2(d) - Y_1(d))$.*

Assumption 1 is a parallel trends assumption, on all the potential outcomes. It requires that $Y_2(d) - Y_1(d)$ be mean independent of the treatments. While this assumption does not restrict treatment effect heterogeneity, it for instance implies that the treatment effect should follow the same evolution over time among movers and stayers: for every $d \neq d'$,

$$\begin{aligned} &E(Y_2(d) - Y_2(d')|M = 1) - E(Y_1(d) - Y_1(d')|M = 1) \\ =&E(Y_2(d) - Y_2(d')|M = 0) - E(Y_1(d) - Y_1(d')|M = 0). \end{aligned}$$

Assumption 2 *(Bounded treatment and Lipschitz potential outcomes)*

1. *$\mathcal{D}_1$ and $\mathcal{D}_2$ are bounded subsets of $\mathbb{R}$.*
2. *For all $(d, d') \in \mathcal{D}_2^2$, there is a constant M such that $|Y_2(d) - Y_2(d')| \leq M|d - d'|$ almost surely.*

Assumption 2 is a technical condition ensuring that all the expectations below are well defined. It requires that the set of values that the period-one and period-two treatments can take be bounded. It also requires that the period-2 potential outcome be a Lipschitz function. This latter requirement will for instance automatically hold if $Y_2(d)$ is differentiable with respect to d and has a bounded derivative.

3 Identification results

We first make the following support assumption.

Assumption 3 *(Support condition) $0 < P(M = 1)$, and $Supp(D_1|M = 1) \subset Supp(D_1|M = 0)$: almost surely, $P(M = 1|D_1) < 1$.*

This is the standard support condition for matching estimators. Note that Assumption 3 implies that $P(M = 0) > 0$: while we assume that the treatments D_1 and D_2 are continuous, in this section we also assume that the treatment is persistent, and thus $D_2 - D_1$ has a mixed distribution with a mass point at zero.

Under Assumption 1-3, one has the following identification result.

Theorem 1 *If Assumptions 1-3 hold,*

$$\begin{aligned}\delta_1 &:= E\left(\frac{Y_2(D_2)-Y_2(D_1)}{D_2-D_1}\middle| M=1\right)\\ &= E\left(\frac{Y_2-Y_1-E(Y_2-Y_1|D_1,M=0)}{D_2-D_1}\middle| M=1\right).\end{aligned}$$

The causal effect δ_1 identified in Theorem 1, is the average, across all movers, of the effect of moving their treatment from its period-one to its period-two value, scaled by the difference between these two values. In other words, δ_1 is the average slope of mover's potential outcome function, from their period-one to their period-two treatment, which we refer to as the average of movers' potential outcome slope (AMPOS).

The AMPOS averages effects of discrete rather than infinitesimal changes in the treatment as in Hoderlein and White (2012), for instance. But if one slightly reinforces Point 2 of Assumption 2 by supposing that $d \mapsto Y_2(d)$ is differentiable on $\mathcal{D}_1 \cup \mathcal{D}_2$, by the mean value theorem,

$$\frac{Y_2(D_2)-Y_2(D_1)}{D_2-D_1} = Y_2'(\widetilde{D})$$

for some $\widetilde{D} \in (\min(D_1,D_2), \max(D_1,D_2))$. Then, δ_1 is an average marginal effect on movers, $\delta_1 = E[Y_2'(\widetilde{D})|M=1]$. The only difference with the usual average marginal effect on movers $E[Y_2'(D_2)|M=1]$ is that it is evaluated at $\widetilde{D}$ instead of D_2.

Intuitively, movers' treatment effects are identified by comparing their outcome evolution to that of stayers with the same period-one treatment. Under our parallel trends assumption, this latter evolution is a valid counterfactual of the outcome evolution that movers would have experienced without that move. Movers' treatment effect $Y_2(D_2)-Y_2(D_1)$ is normalized by D_2-D_1, to measure by how much a one-unit increase in treatment changes the outcome on average.

In practice, the estimand in Theorem 1 can be estimated in two steps. One first estimates a non-parametric regression of Y_2-Y_1 on D_1 among stayers. Then, one computes the average of

$$\frac{Y_2-Y_1-\widehat{E}(Y_2-Y_1|D_1,M=0)}{D_2-D_1}$$

among movers, where $\widehat{E}(Y_2-Y_1|D_1,M=0)$ is the predicted value from the non-parametric regression.

We now consider a second strategy to identify other causal effects, under the following support conditions.

Assumption 4 *(Support condition on subpopulations)*

1. $0 < P(M_i=1)$, *and* $Supp(D_1|M_i=1) \subset Supp(D_1|M=0)$*: almost surely,* $0 < P(M_i=1|D_1)$ *implies that* $0 < P(M=0|D_1)$.

2. $0 < P(M_d = 1)$*, and* $Supp(D_1|M_d = 1) \subset Supp(D_1|M = 0)$*: almost surely,* $0 < P(M_d = 1|D_1)$ *implies that* $0 < P(M = 0|D_1)$*.*

Point 1 (resp. 2) requires that a similar support condition as Assumption 3 holds in the subpopulation of movers whose treatment increases (resp. decreases) and stayers.

Theorem 2 *1. If Assumptions 1-2 and Point 1 of Assumption 4 holds,*

$$\begin{aligned}\delta_{2i} :=& \frac{E\left(Y_2(D_2) - Y_2(D_1)|M_i = 1\right)}{E(D_2 - D_1|M_i = 1)} \\ =& \frac{E\left(Y_2 - Y_1 - E(Y_2 - Y_1|D_1, M = 0)|M_i = 1\right)}{E(D_2 - D_1|M_i = 1)} \qquad (1) \\ =& \frac{E\left(Y_2 - Y_1|M_i = 1\right) - E\left((Y_2 - Y_1)\frac{P(M_i=1|D_1)}{P(M=0|D_1)}\frac{P(M=0)}{P(M_i=1)}\Big|M = 0\right)}{E(D_2 - D_1|M_i = 1)}. \qquad (2)\end{aligned}$$

2. If Assumptions 1-2 and Point 2 of Assumption 4 holds,

$$\begin{aligned}\delta_{2d} :=& \frac{E\left(Y_2(D_1) - Y_2(D_2)|M_d = 1\right)}{E(D_1 - D_2|M_d = 1)} \\ =& \frac{E\left(Y_2 - Y_1 - E(Y_2 - Y_1|D_1, M = 0)|M_d = 1\right)}{E(D_2 - D_1|M_d = 1)} \qquad (3) \\ =& \frac{E\left(Y_2 - Y_1|M_d = 1\right) - E\left((Y_2 - Y_1)\frac{P(M_d=1|D_1)}{P(M=0|D_1)}\frac{P(M=0)}{P(M_d=1)}\Big|M = 0\right)}{E(D_2 - D_1|M_d = 1)}. \qquad (4)\end{aligned}$$

The causal effects identified in Points 1 and 2 of Theorem 2 respectively apply to movers whose treatment increases and decreases. It directly follows from Theorem 2 that

$$\delta_2 :=P(M_i = 1|M = 1)\delta_{2i} + P(M_d = 1|M = 1)\delta_{2d} \qquad (5)$$

is identified. δ_2 applies to all movers, like δ_1.

Even then, there remains a difference between δ_2 and δ_1. One has

$$\delta_{2i} = E\left(\frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}\frac{D_2 - D_1}{E(D_2 - D_1|M_i = 1)}\Big|M_i = 1\right).$$

Accordingly, δ_{2i} is a weighted average, across movers whose treatment increases, of the slope of each mover's potential outcome function from its period-one to its period-two treatment, where units receive a weight proportional to the absolute value of their treatment change from period one to two. A similar interpretation applies to δ_{2d}, among the population of movers whose treatment decreases. Accordingly, we refer to δ_2 as the weighted average of movers' potential outcome slope (WAMPOS). δ_2 may differ from δ_1, if movers' potential outcome slopes are correlated with the magnitude of their treatment change from period one to two.

δ_2 may look like a less natural causal effect than δ_1. However, the estimator based on Theorem 1 could be much noisier than the estimators based on Theorem 2, if there are movers whose treatment barely changes from period one to two. Specifically, Graham and Powell (2012) show that the semiparametric efficiency bound for a parameter very similar to δ_1 is 0 when there exists an interval A such that

$$\{0\} \subsetneq A \subset \text{Supp}(D_2 - D_1), \tag{6}$$

meaning that $D_2 - D_1$, the treatment evolution from period one to two, can take values arbitrarily close to 0. This result implies that there exists no regular root-n consistent estimator of such a parameter. Intuitively, one needs to trim observations for which $D_2 - D_1$ is too close to 0, because for such observations, the variance of

$$\frac{Y_2 - Y_1 - \widehat{E}(Y_2 - Y_1 | D_1, M = 0)}{D_2 - D_1}$$

becomes very large. This trimming results in a non-negligible bias, and as in standard nonparametric estimation (e.g., of a density), in a lower than root-n rate of convergence. On the other hand, even if (6) holds, one can expect the semiparametric efficiency bound to be positive for δ_2, meaning that we can estimate it at the standard root-n rate. If (6) does not hold, meaning that the treatment evolution cannot take values arbitrarily close to 0 (though it can be exactly equal to 0), one may still expect that if $\inf \text{Supp}(|D_2 - D_1|)$ is small, an estimator of δ_1 will be noisier than that of δ_2.

Point 1 of Theorem 2 shows that δ_{2i} is identified by two estimands, a regression-based and a propensity-score-based estimand. Though we do not formally show it here, δ_{2i} could also be identified by a doubly-robust estimand.

The regression based estimand in Equation (1) compares the outcome evolution of movers whose treatment increases to that of stayers with the same period-one treatment, averages such comparisons, and then normalizes this average by the average treatment evolution among movers whose treatment increases. As explained in the introduction, this estimand corresponds, after some relabelling, to the W^X_{DID} estimand proposed in Theorem S4 of the supplement of de Chaisemartin and D'Haultfœuille (2018). In Equation (1), the M_i indicator and the period-one treatment D_1 respectively play the role of the treatment-group indicator and of the covariates in the W^X_{DID} estimand. Accordingly, an estimator of δ_{2i} can readily be computed using the `fuzzy_did` Stata command (see de Chaisemartin et al., 2019), which computes the $\widehat{W}^X_{DID}$ estimator of W^X_{DID} proposed by de Chaisemartin and D'Haultfœuille (2018).

The propensity-score based estimand in Equation (2) compares the average outcome evolution of movers to that of stayers, reweighting stayers to ensure that they have the same distribution of the period-1 treatment as movers. Then, it normalizes this comparison by the average treatment evolution among movers whose treatment increases. The numerator of the estimand in

Equation (2) corresponds, after some relabelling, to the propensity-score reweighting DID estimand in Equation (10) of Abadie (2005). In Equation (2), the M_i indicator and the period-one treatment D_1 respectively play the role of the treatment and of the covariates in Abadie (2005). Accordingly, an estimator of that numerator can be computed using the `absdid` Stata command (see Houngbedji, 2016). Similar interpretations and discussions apply to the two estimands in Point 2 of Theorem 2.

While in this paper we focus on continuous treatments, our results can also be applied to discrete treatments. In Section 4 of their Web Appendix, de Chaisemartin and D'Haultfœuille (2020*b*) already propose a DID estimator of the effect of a discrete treatment. The plug-in estimator of δ_2 one can form following Theorem 2 and using simples averages to estimate the non-parametric regressions or the propensity scores is numerically equivalent to the estimator therein. This paper still makes two contributions relative to de Chaisemartin and D'Haultfœuille (2020*b*) when the treatment is discrete. First, the estimator based on Theorem 1 was not proposed therein. Second, with a discrete treatment taking a large number of values, the estimator in de Chaisemartin and D'Haultfœuille (2020*b*) may not be applicable as it requires finding movers and stayers with the exact same period-one treatment, which may not always be feasible. Instead, one can follow Theorem 2, using a parametric model, to estimate the regressions or the propensity scores entering the estimands in that theorem.

4 Extensions

For brevity, in this section we only consider extensions of Theorem 1, and of the results in Theorem 2 involving propensity-score reweighting DID estimands. Results in Theorem 2 involving non-parametric regressions can be extended following the same steps as those used to extend Theorem 1.

4.1 No stayers

So far we have assumed that $P(M = 0) > 0$, meaning that there are units whose treatment does not change over time. We now show that when this is not the case, Theorem 1 above readily extends provided that, roughly speaking, quasi-stayers exist. Specifically, let $M_\delta = 1\{|D_2 - D_1| > \delta\}$. Quasi-stayers are such that $M_\delta = 0$ for some small $\delta > 0$. We consider the following support condition.

Assumption 5 *(Support condition without stayers)* $0 < P(M = 1)$ *and for all* $\delta > 0$, $Supp(D_1|M_\delta = 1) \subset Supp(D_1|M_\delta = 0)$*: almost surely,* $P(M_\delta = 1|D_1) < 1$.

Because $M_\delta \leq M$ for all $\delta > 0$, Assumption 5 is weaker than Assumption 3. In particular, it does not require that $P(M = 0) > 0$.

Theorem 3 *If Assumptions 1-2 and Assumption 5 holds,*

$$\delta_1 = E\left(\frac{Y_2 - Y_1 - \lim_{\delta \to 0} E(Y_2 - Y_1 | D_1, M_\delta = 0)}{D_2 - D_1}\middle| M = 1\right).$$

While the identification results in Theorems 1 and 3 are almost identical, the estimators one would form will depend on whether there are stayers or not in the sample. With stayers, one will estimate $E(Y_2 - Y_1 | D_1 = d_1, M = 0)$ in Theorem 3 by the (potentially weighted) average of $Y_2 - Y_1$ among stayers with $D_1 \in [d_1 - h, d_1 + h]$, for some bandwidth h. Without stayers, one will estimate $\lim_{\delta \to 0} E(Y_2 - Y_1 | D_1 = d_1, M_\delta = 0)$ in Theorem 1 by the (potentially weighted) average of $Y_2 - Y_1$ among quasi-stayers with $D_1 \in [d_1 - h, d_1 + h], D_2 - D_1 \in [-\delta, \delta]$, for two bandwidths h and δ. The idea of using quasi stayers as controls is similar to that in Graham and Powell (2012). We conjecture that the estimator of δ_1 based on Theorem 3 will converge at a slower rate than that based on Theorem 1, that requires stayers. This is because (i) the estimator of $\lim_{\delta \to 0} E(Y_2 - Y_1 | D_1 = d_1, M_\delta = 0)$ converges at a slower rate than that of $E(Y_2 - Y_1 | D_1 = d_1, M = 0)$; (ii) as a result of trimming, the rate of convergence of this first-step estimator has an effect on the final estimator of δ_1. We refer to Graham and Powell (2012) for more details on this last point.

Theorem 2 can also be extended to the case without stayers. For any $\delta \geq 0$, let $M_{i,\delta} = 1\{D_2 - D_1 > \delta\}$ and $M_{d,\delta} = 1\{D_2 - D_1 < -\delta\}$ respectively be indicators for units whose treatment increases and decreases by more than δ from period one to two.

Assumption 6 *(Regularity and support and conditions for the propensity score estimand without stayers)*

1. $\delta \mapsto P(D_2 - D_1 \leq \delta)$ *is continuous at* 0.
2. *There is an interval* $I_+ \subseteq \mathbb{R}_+$ *containing zero and not reduced to a point such that for all* $\delta \in I_i$, $0 < P(M_{i,\delta} = 1)$, *and for all* $\delta \neq 0$, *almost surely* $0 < P(M_{i,\delta} = 1 | D_1) \Rightarrow 0 < P(M_\delta = 0 | D_1)$.
3. *There is an interval* $I_- \subseteq \mathbb{R}_+$ *containing zero and not reduced to a point such that for all* $\delta \in I_-$, $0 < P(M_{d,\delta} = 1)$, *and for all* $\delta \neq 0$, *almost surely* $0 < P(M_{d,\delta} = 1 | D_1) \Rightarrow 0 < P(M_\delta = 0 | D_1)$.

The first point of Assumption 6 assumes that $D_2 - D_1$ does not have a mass point at 0, i.e. that there are no stayers. The second (resp. third) point of Assumption 6 requires that for values of δ close to 0, the support of D_1 be the same among quasi-stayers whose treatment changes by less than δ and movers whose treatment increases (resp. decreases) by more than δ.

Theorem 4 *1. If Assumptions 1-2 and Points 1 and 2 of Assumption 6 hold,*

$$\delta_{2i} = \lim_{\delta \to 0} \frac{E\left(Y_2 - Y_1 | M_{i,\delta} = 1\right) - E\left((Y_2 - Y_1)\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta=0|D_1)}\frac{P(M_\delta=0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\right)}{E(D_2 - D_1 | M_{i,\delta} = 1)}.$$

2. If Assumptions 1-2 and Points 1 and 3 of Assumption 6 hold,

$$\delta_{2d} = \lim_{\delta \to 0} \frac{E\left(Y_2 - Y_1 | M_{d,\delta} = 1\right) - E\left((Y_2 - Y_1)\frac{P(M_{d,\delta}=1|D_1)}{P(M_\delta=0|D_1)}\frac{P(M_\delta=0)}{P(M_{d,\delta}=1)}\Big| M_\delta = 0\right)}{E(D_2 - D_1 | M_{d,\delta} = 1)}.$$

Theorem 4 shows that without stayers, the same causal effects as in Theorem 2 are still identified, by limits of propensity score matching estimands similar to those in Theorem 2, except that stayers are replaced by quasi-stayers whose treatment increases by less than δ, while movers are replaced by movers whose treatment increases by more than δ.

4.2 More than two time periods

In this section, we assume the representative unit is observed at $T > 2$ time periods. Let $(D_1, ..., D_T)$ denote the unit's treatments and $\mathcal{D}_t = \text{Supp}(D_t)$ for all $t \in \{1, ..., T\}$. For any $t \in \{1, ..., T\}$, and for any $d \in \cup_{t=1}^{T} \mathcal{D}_t$ let $Y_t(d)$ denote the unit's potential outcome at period t with treatment d. Finally, let Y_t denote her observed outcome at t. For any $t \in \{2, ..., T\}$, let $M_t = 1\{D_t \neq D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment moves from period $t-1$ to t. Let also $M_{i,t} = 1\{D_t > D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment increases from period $t-1$ to t, and let $M_{d,t} = 1\{D_t < D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment decreases.

We make the following assumptions, which generalize Assumptions 1 and 3 to settings with more than two time periods.

Assumption 7 *(Parallel trends) For all $t \geq 2$, for all $d \in \mathcal{D}_{t-1} \cup \mathcal{D}_t$, almost surely, $E(Y_t(d) - Y_{t-1}(d)|D_{t-1}, D_t) = E(Y_t(d) - Y_{t-1}(d))$.*

Assumption 8 *(Support conditions for three or more periods) For all $t \geq 2$, $0 < P(M_t = 1)$, and the support of $D_{t-1}|M_t = 1$ is included in that of $D_{t-1}|M_t = 0$: almost surely, $P(M_t = 1|D_{t-1}) < 1$.*

Assumption 9 *(Support conditions on subpopulations for three or more periods) For all $t \geq 2$,*

1. *$0 < P(M_{i,t} = 1)$, and the support of $D_{t-1}|M_{i,t} = 1$ is included in that of $D_{t-1}|M_t = 0$: almost surely, $0 < P(M_{i,t} = 1|D_{t-1}) \Rightarrow 0 < P(M_t = 0|D_{t-1})$.*

2. *$0 < P(M_{d,t} = 1)$, and the support of $D_{t-1}|M_{d,t} = 1$ is included in that of $D_{t-1}|M_t = 0$: almost surely, $0 < P(M_{d,t} = 1|D_{t-1}) \Rightarrow 0 < P(M_t = 0|D_{t-1})$.*

Theorems 5 and 6 below generalize Theorems 1 and 2 to settings with more than two time periods. As above, we define

$$\delta_{1t} = E\left(\frac{Y_t(D_t) - Y_t(D_{t-1})}{D_t - D_{t-1}}\middle| M_t = 1\right),$$
$$\delta_{2it} = \frac{E\left(Y_t(D_t) - Y_t(D_{t-1})|M_{i,t} = 1\right)}{E(D_t - D_{t-1}|M_{i,t} = 1)},$$
$$\delta_{2dt} = \frac{E\left(Y_t(D_{t-1}) - Y_t(D_t)|M_{d,t} = 1\right)}{E(D_{t-1} - D_t|M_{d,t} = 1)}.$$

Theorem 5 *If Assumption 7 and Point 1 of Assumption 8 hold,*

$$\sum_{t=2}^{T} \frac{P(M_t = 1)}{\sum_{k=2}^{T} P(M_k = 1)}\delta_{1t} = \sum_{t=2}^{T} \frac{P(M_t = 1)}{\sum_{k=2}^{T} P(M_k = 1)} E\left[\frac{Y_t - Y_{t-1} - E(Y_t - Y_{t-1}|D_{t-1}, M_t = 0)}{D_t - D_{t-1}}\middle| M_t = 1\right].$$

Theorem 6 *If Assumption 7 and Points 2 and 3 of Assumption 8 hold,*

$$\begin{aligned}
&\sum_{t=2}^{T} \frac{P(M_{i,t} = 1)}{\sum_{k=2}^{T} P(M_k = 1)}\delta_{2it} + \frac{P(M_{d,t} = 1)}{\sum_{k=2}^{T} P(M_k = 1)}\delta_{2dt} \\
=&\sum_{t=2}^{T}\left[\frac{P(M_{i,t} = 1)}{\sum_{k=2}^{T} P(M_k = 1)} \frac{E\left(Y_t - Y_{t-1}|M_{i,t} = 1\right) - E\left((Y_t - Y_{t-1})\frac{P(M_{i,t}=1|D_{t-1})}{P(M_t=0|D_{t-1})}\frac{P(M_t=0)}{P(M_{i,t}=1)}\middle| M_t = 0\right)}{E(D_t - D_{t-1}|M_{i,t} = 1)}\right. \\
&\left. + \frac{P(M_{d,t} = 1)}{\sum_{k=2}^{T} P(M_k = 1)} \frac{E\left(Y_t - Y_{t-1}|M_{d,t} = 1\right) - E\left((Y_t - Y_{t-1})\frac{P(M_{d,t}=1|D_{t-1})}{P(M_t=0|D_{t-1})}\frac{P(M_t=0)}{P(M_{d,t}=1)}\middle| M_t = 0\right)}{E(D_{t-1} - D_t|M_{d,t} = 1)}\right].
\end{aligned}$$

The estimator based on Theorem 5 relies on $T-1$ non-parametric regressions of $Y_t - Y_{t-1}$ on D_{t-1}, among units with $M_t = 0$, for $t \geq 2$. The estimator based on Theorem 6 relies on $3(T-1)$ propensity score estimations, of $M_{i,t}$, $M_{d,t}$, and $1 - M_t$ on D_{t-1}, for $t \geq 2$.

In this section, we assume that there are stayers, meaning that the treatment is persistent over time. Then, it may be the case that among the period-t movers, some keep the same treatment till period $t + \ell$, for $\ell > 0$. Below, we show that their effect of having switched from D_{t-1} to D_t can be identified till $t + \ell$, under a mild strengthening of Assumptions 7 and 8. For brevity, we do not propose propensity-score reweighting estimands of those long-run effects, but proposing such estimands would be straightforward.

Assumption 10 *(Parallel trends, to estimate long-run effects) For all $t \geq 2$, for all $d \in \cup_{k=t-1}^{T} \mathcal{D}_{t-1}$, almost surely, $E(Y_t(d) - Y_{t-1}(d)|D_{t-1}, D_t, ..., D_T) = E(Y_t(d) - Y_{t-1}(d))$.*

Assumption 11 *(Support conditions, to estimate long-run effects) For all $t \geq 2$ and $\ell \geq 1$ such that $t + \ell \leq T$, $0 < P\left(M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0\right)$, and the support of $D_{t-1}|M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0$ is included in that of $D_{t-1}|M_t = 0, \sum_{k=1}^{\ell} M_{t+k} = 0$:*

$$\textit{almost surely, } 0 < P\left(M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0|D_{t-1}\right) \Rightarrow 0 < P\left(M_t = 0, \sum_{k=1}^{\ell} M_{t+k} = 0|D_{t-1}\right).$$

The support condition above may fail for large values of ℓ. In that case, the theorems below only hold for values of ℓ for which those support conditions hold.

We now define, for every t, the long-run effects among the period-t movers who keep the same treatment for ℓ periods after t. For every $t \geq 2$ and $\ell \geq 1$ such that $t + \ell \leq T$, let

$$\delta_{1t:t+\ell} = E \quad \frac{Y_{t+\ell}(D_t) - Y_{t+\ell}(D_{t-1})}{D_t - D_{t-1}} \Bigg| M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0 \Bigg).$$

Theorem 7 *If Assumption 10 and Point 1 of Assumption 11 hold, then for every $\ell \in \{1, ..., T-2\}$*

$$\begin{aligned}
&\sum_{t=2}^{T-\ell} \frac{P\left(M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0\right)}{\sum_{j=2}^{T-\ell} P\left(M_j = 1, \sum_{k=1}^{\ell} M_{j+k} = 0\right)} \delta_{1t:t+\ell} \\
&= \sum_{t=2}^{T-\ell} \Bigg[\frac{P\left(M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0\right)}{\sum_{j=2}^{T-\ell} P\left(M_j = 1, \sum_{k=1}^{\ell} M_{j+k} = 0\right)} \\
&\times E\left(\frac{Y_{t+\ell} - Y_{t-1} - E\left(Y_{t+\ell} - Y_{t-1} | D_{t-1}, M_t = 0, \sum_{k=1}^{\ell} M_{t+k} = 0\right)}{D_t - D_{t-1}} \Bigg| M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0 \right) \Bigg].
\end{aligned}$$

4.3 Allowing for dynamic effects

In this subsection, we allow for dynamic effects. The results below generalize those in de Chaisemartin and D'Haultfœuille (2020*a*), who allow for dynamic effects, but require that the treatment be either discrete, see Section 1.3 of their Web Appendix, or continuous but following a staggered adoption design, see Section 1.4 of their Web Appendix. Below, we allow for a continuous treatment that may not follow a staggered adoption design. We still require that the representative unit's treatment can never get lower than her period-one treatment:

Assumption 12 *(Lowest treatment at period one) For all t, $D_t \geq D_1$.*

It may be the case that for some units, the treatment is at its lowest at period one, but for other units the treatment is at its highest at period one. In that case, one can split the sample in two, and compute the estimators based on the results below in the first subsample, and the negative of those estimators in the second subsample. There may also be some units that have a treatment higher than their period-one treatment at some time periods, but a lower treatment at other time periods. One may have to discard such units, as the dynamic treatment effects of those units may conflate together effects of increases and decreases of the treatment, and may not be interpretable, see Section 1.3 of the Web Appendix of de Chaisemartin and D'Haultfœuille (2020*a*) for a discussion.

For all $\boldsymbol{d} \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, let $Y_t(\boldsymbol{d})$ denote the potential outcome of the representative unit at period t, if her treatments from period one to T are equal to $\boldsymbol{d}$. This dynamic potential outcome framework is similar to that in Robins (1986). It allows for the possibility that the outcome at time t be affected by past and future treatments. Let $\boldsymbol{D} = (D_1, ..., D_T)$ be a $1 \times T$ vector stacking the representative unit's treatments from period one to T.

Assumption 13 *(No Anticipation) For all $\boldsymbol{d} \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, $Y_t(\boldsymbol{d}) = Y_t(d_1, ..., d_t)$.*

Assumption 13 requires that the current outcome do not depend on future treatments, the so-called no-anticipation hypothesis. Abbring and Van den Berg (2003) have discussed that assumption in the context of duration models, and Malani and Reif (2015), Botosaru and Gutierrez (2018), and Sun and Abraham (2020) have discussed it in the context of DID models.

Let $F = \min\{t : D_t \neq D_{t-1}\}$ denote the first date at which the representative unit's treatment changes, with the convention that $F = T + 1$ if her treatment never changes. For all $d \in \mathcal{D}_1$, we let $\boldsymbol{d} = (d, ..., d)$ denote a $1 \times T$ vector with coordinates equal to d. We also let $\boldsymbol{D}_1 = (D_1, ..., D_1)$ denote a $1 \times T$ vector with coordinates equal to D_1, the unit's period-one treatment. We make the following assumptions, which generalize Assumptions 1 and 3 to settings with more than two time periods and dynamic effects.

Assumption 14 *(Parallel trends, allowing for dynamic effects) For all $t \geq 2$, for all $d \in \mathcal{D}_1$, almost surely, $E(Y_t(\boldsymbol{d}) - Y_{t-1}(\boldsymbol{d})|\boldsymbol{D}) = E(Y_t(\boldsymbol{d}) - Y_{t-1}(\boldsymbol{d}))$.*

Assumption 15 *(Support condition, allowing for dynamic effects) For all $t \geq 2$, $0 < P(F = t)$, and for all $t' > t$, the support of $D_1|F = t$ is included in that of $D_1|F > t'$: almost surely, $0 < P(F = t|D_1) \Rightarrow 0 < P(F > t'|D_1)$.*

Assumption 15 requires that the probability that the representative unit never changes its treatment be strictly positive. When that condition fails, results below still hold, till the last period where the probability of having never changed treatment is still strictly positive.

Theorem 8 below generalizes Theorems 5 and 6 to allow for dynamic effects.

Theorem 8 *If Assumptions 13, 14 and 15 hold, for any $\ell \in \{0, ..., T-2\}$,*

$$\begin{aligned}
\delta_{+,\ell} &:= \sum_{t=\ell+2}^{T} \frac{P(F = t-\ell)}{\sum_{k=\ell+2}^{T} P(F = k-\ell)} E\left(Y_t(\boldsymbol{D}) - Y_t(\boldsymbol{D}_1)|F = t-\ell\right) \\
&= \sum_{t=\ell+2}^{T} \frac{P(F = t-\ell)}{\sum_{k=\ell+2}^{T} P(F = k-\ell)} E\left(Y_t - Y_{t-\ell-1} - E(Y_t - Y_{t-\ell-1}|D_1, F > t)|F = t-\ell\right) \quad (7) \\
&= \sum_{t=\ell+2}^{T} \frac{P(F = t-\ell)}{\sum_{k=\ell+2}^{T} P(F = k-\ell)} \left(E\left(Y_t - Y_{t-\ell-1}|F = t-\ell\right)\right. \\
&\quad \left. - E\left((Y_t - Y_{t-\ell-1})\frac{P(F = t-\ell|D_1)}{P(F > t|D_1)}\frac{P(F > t)}{P(F = t-\ell)}\middle| F > t\right)\right). \quad (8)
\end{aligned}$$

The causal effects $\delta_{+,\ell}$ identified in Theorem 8 are reduced-form effects, comparing in period t the actual outcome of units whose treatment changed for the first time ℓ periods ago to the counterfactual outcome they would have obtained if they had instead kept their period-one treatment from period one to t. Assumption 12 ensures that those effects can be interpreted as effects of increasing the treatment, but they may aggregate together the effects of many different treatment trajectories. For instance, if $T = 4$, some units may be such that $(D_1 = 1, D_2 = 1, D_3 = 2, D_4 = 3)$, while other units may be such that $(D_1 = 2, D_2 = 2, D_3 = 3, D_4 = 2)$. Then, $\delta_{+,\ell}$ aggregates together $Y(1,1,2,3) - Y(1,1,1,1)$ for the first set of units, and $Y(2,2,3,2) - Y(2,2,2,2)$ for the second set of units. This complicates the interpretation of $\delta_{+,\ell}$. For instance, those parameters cannot be interpreted as averages of the effect of increasing the treatment by one unit on the outcome.

However, following Lemma 1 in de Chaisemartin and D'Haultfœuille (2020a), the weighted sum of those reduced-form effects in Equation (9) below has a clear economic interpretation. Let

$$w_\ell = \frac{\sum_{k=\ell+2}^{T} P(F = k-\ell)}{\sum_{j=0}^{T-2}\sum_{k=j+2}^{T} P(F = k-j)},$$

$$\delta^D_{+,\ell} := \sum_{t=\ell+2}^{T} \frac{P(F = t-\ell)}{\sum_{k=\ell+2}^{T} P(F = k-\ell)} E\left(D_t - D_1 | F = t-\ell\right),$$

and

$$\delta_+ := \frac{\sum_{\ell=0}^{T-2} w_\ell \delta_{+,\ell}}{\sum_{\ell=0}^{T-2} w_\ell \delta^D_{+,\ell}}. \tag{9}$$

δ_+ starts by averaging across ℓ the reduced-form effects $\delta_{+,\ell}$, with weights proportional to the probability that before the end of the panel, a unit is observed ℓ periods after its first treatment switch. Then, it divides that average by the difference between the average actual treatments received by movers and the treatments they would have received if they had kept all along their period-one treatment. As discussed in de Chaisemartin and D'Haultfœuille (2020a), this ratio may then be used in a cost-benefit analysis comparing the actual treatments to the status quo scenario where movers would have kept their period-one treatment all along. It can also be interpreted as some average of the effect of increasing the treatment by one unit on the outcome.

Equation (7) shows that $\delta_{+,\ell}$ is identified by a regression-based estimand, while Equation (8) shows that it is also identified by a propensity-score-reweighting-based estimand. One could show that $\delta_{+,\ell}$ is also identified by a doubly-robust estimand.

The reduced-form estimands we propose in Equation (7) (resp. Equation (8)) are equal to a weighted average of the estimands proposed in Equation (2.6) (resp. Equation (2.5)) of Callaway and Sant'Anna (2020) to identify the instantaneous and dynamic effects of a binary and staggered treatment. In the definition of $\delta_{+,\ell}$, the indicator $F = t - \ell$ for having changed treatment for

the first time at $t-\ell$ plays the role of the indicator for having been treated for the first time at $t-\ell$ in Callaway and Sant'Anna (2020), the indicator $F > t$ for having never changed treatment at t plays the role of the indicator for not having been treated yet at t in Callaway and Sant'Anna (2020), and D_1 plays the role of the covariates in Callaway and Sant'Anna (2020). This implies that estimators of $\delta_{+,\ell}$ attached to the estimands we propose in Equations (7) and (8) can be computed using, say, the `csdid` Stata command (see Rios-Avila et al., 2021). Then, an estimator of the more interpretable δ_+ parameter can be computed by replacing the $\delta_{+,\ell}$s by their estimators in Equation (9), and by replacing all the other population quantities by their sample equivalents.

5 Future work

In future work, we will formally study the asymptotic properties of the estimators based on our identification results. We will start by considering the case with some stayers and no quasi-stayers, where we anticipate that estimators of the AMPOS and WAMPOS will both converge at the parametric rate. We will then consider the case with some stayers and quasi-stayers, where we anticipate that only the estimators of the WAMPOS will converge at the parametric rate. Finally, we will then consider the case with no stayers and some quasi-stayers, where we anticipate that estimators of the AMPOS and WAMPOS will both converge at a non-parametric rate. We will also use our results to revisit Fajgelbaum et al. (2020) and Deschênes and Greenstone (2007).

6 Proofs

6.1 Theorem 1

Fix $d_1 \in \text{Supp}(D_1|M=1)$. By Assumption 3, $d_1 \in \text{Supp}(D_1|M=0)$. Thus, $E(Y_2 - Y_1|D_1 = d_1, M=0)$ is well-defined. Next, for every $d_2 \neq d_1$ such that $(d_1, d_2) \in \text{Supp}(D_1, D_2|M=1)$,

$$\begin{aligned} E(Y_2 - Y_1|D_1 = d_1, M=0) =& E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_1) \\ =& E(Y_2(d_1) - Y_1(d_1)) \\ =& E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_2, M=1). \end{aligned} \tag{10}$$

The first equality follows by definition of M. The second and third equalities follow from Assumption 1. Then,

$$\begin{aligned}
&E\left(\frac{Y_2 - Y_1 - E(Y_2 - Y_1|D_1, M=0)}{D_2 - D_1}\middle| M=1\right)\\
=&E\left(\frac{E(Y_2 - Y_1|D_1, D_2, M=1) - E(Y_2 - Y_1|D_1, M=0)}{D_2 - D_1}\middle| M=1\right)\\
=&E\left(\frac{E(Y_2(D_2) - Y_1(D_1)|D_1, D_2, M=1) - E(Y_2(D_1) - Y_1(D_1)|D_1, D_2, M=1)}{D_2 - D_1}\middle| M=1\right)\\
=&E\left(\frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}\middle| M=1\right),
\end{aligned}$$

where the third equality follows from (10), and where the expectation in the last equality is well defined by Assumption 2 □

6.2 Theorem 2

We only prove the first point, the proof of the second point is symmetric. The proof of Equation (1) is similar to the proof of Theorem 1 so it is omitted. We now prove Equation 2. Using the same reasoning as above, for every d_1 in the support of $D_1|M_i = 1$,

$$E(Y_2 - Y_1|D_1 = d_1, M = 0) = E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, M_i = 1).$$

Hence, by repeated use of the law of iterated expectation,

$$\begin{aligned}
&E\left((Y_2 - Y_1)\frac{P(M_i=1|D_1)}{P(M=0|D_1)}\frac{P(M=0)}{P(M_i=1)}\middle| M=0\right)\\
=&E\left(E[Y_2(D_1) - Y_1(D_1)|D_1, M_i=1]\frac{P(M_i=1|D_1)}{P(M=0|D_1)}\frac{P(M=0)}{P(M_i=1)}\middle| M=0\right)\\
=&E\left(E[Y_2(D_1) - Y_1(D_1)|D_1, M_i=1]\frac{P(M_i=1|D_1)}{P(M=0|D_1)}\frac{1-S}{P(M_i=1)}\right)\\
=&E\left(E[Y_2(D_1) - Y_1(D_1)|D_1, M_i=1]\frac{P(M_i=1|D_1)}{P(M_i=1)}\right)\\
=&E\left(E[Y_2(D_1) - Y_1(D_1)|D_1, M_i=1]\frac{S}{P(M_i=1)}\right)\\
=&E\left(Y_2(D_1) - Y_1(D_1)|M_i=1\right).
\end{aligned}$$

The result follows after some algebra, once noted that all expectations are well defined by Assumption 2 □

6.3 Theorem 3

Fix $\delta > 0$. By Assumption 5, $P(M_\delta = 0|D_1) > 0$ almost surely. Thus, the conditional expectation $E(Y_2 - Y_1|D_1, M_\delta = 0)$ is well-defined. Moreover,

$$E(Y_2 - Y_1|D_1, M_\delta = 0) = E(Y_2(D_2) - Y_2(D_1)|D_1, M_\delta = 0) + E(Y_2(D_1) - Y_1(D_1)|D_1, M_\delta = 0). \tag{11}$$

Now, by Jensen's inequality and Point 2 of Assumption 2,

$$\begin{aligned} |E(Y_2(D_2) - Y_2(D_1)|D_1, M_\delta = 0)| \leq & E(|Y_2(D_2) - Y_2(D_1)| \ |D_1, M_\delta = 0) \\ \leq & ME(|D_2 - D_1| \ |D_1, M_\delta = 0) \\ \leq & M\delta. \end{aligned} \tag{12}$$

Next, by Assumption 1,

$$\begin{aligned} E(Y_2(D_1) - Y_1(D_1)|D_1, M_\delta = 0) &= E(Y_2(D_1) - Y_1(D_1)) \\ &= E(Y_2(D_1) - Y_1(D_1)|D_1, D_2, M = 1). \end{aligned}$$

Combined with (11)-(12), this yields

$$\lim_{\delta \to 0} E(Y_2 - Y_1|D_1, M_\delta = 0) = E(Y_2(D_1) - Y_1(D_1)|D_1, D_2, M = 1).$$

The remainder of the proof is similar to that of Theorem 1 □

6.4 Theorem 4

We only prove the first point, the proof of the second point is symmetric. We consider an arbitrary $\delta > 0$ in I_+. Point 1 of Assumption 2 ensures that there is a constant B such that $|D_2 - D_1| \leq B$ almost surely. Then,

$$\begin{aligned} & |E(Y_2(D_2) - Y_2(D_1)|M_i = 1) - E(Y_2(D_2) - Y_2(D_1)|M_{i,\delta} = 1)| \\ = & \frac{P(D_2 - D_1 \in (0, \delta])}{P(D_2 - D_1 > 0)} |E(Y_2(D_2) - Y_2(D_1)|D_2 - D_1 \in (0, \delta]) - E(Y_2(D_2) - Y_2(D_1)|D_2 - D_1 > \delta)| \\ \leq & \frac{P(D_2 - D_1 \leq \delta) - P(D_2 - D_1 \leq 0)}{P(D_2 - D_1 > 0)} M(\delta + B), \end{aligned}$$

where the equality follows from the law of iterated expectations, and the inequality follows from the triangle inequality, Jensen's inequality, and the second and first points of Assumption 2. The previous display and Point 1 of Assumption 6 imply

$$\lim_{\delta \to 0} E(Y_2(D_2) - Y_2(D_1)|M_{i,\delta} = 1) = E(Y_2(D_2) - Y_2(D_1)|M_i = 1). \tag{13}$$

Similarly, one can show that

$$\lim_{\delta \to 0} E\left(D_2 - D_1 | M_{i,\delta} = 1\right) = E\left(D_2 - D_1 | M_i = 1\right). \tag{14}$$

Then, one has

$$\begin{aligned}
& E \quad (Y_2 - Y_1)\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big) \\
=& E \quad (Y_2(D_2) - Y_2(D_1))\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big) \\
+& E \quad (Y_2(D_1) - Y_1(D_1))\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big).
\end{aligned} \tag{15}$$

Then,

$$\begin{aligned}
& \left| E \quad (Y_2(D_2) - Y_2(D_1))\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big)\right| \\
\leq & M E \quad |D_2 - D_1|\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big) \\
\leq & M\delta E \quad \frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big) \\
= & M\delta.
\end{aligned} \tag{16}$$

The first inequality follows from Point 2 of Assumption 2, the second one follows from the definition of M_δ. The equality follows from similar algebra as in the proof of Theorem 2. Then, using the same reasoning as in the proof of Theorem 2, one can show that

$$\begin{aligned}
& E \quad (Y_2(D_1) - Y_1(D_1))\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big) \\
= & E\left(Y_2(D_1) - Y_1(D_1)|M_{i,\delta} = 1\right).
\end{aligned} \tag{17}$$

Then, it follows from Equations (13), (15), (16), and (17) that

$$\begin{aligned}
& E\left(Y_2(D_2) - Y_2(D_1)|M_i = 1\right) \\
= & \lim_{\delta \to 0} E\left(Y_2 - Y_1|M_{i,\delta} = 1\right) - E \quad (Y_2 - Y_1)\frac{P(M_{i,\delta}=1|D_1)}{P(M_\delta = 0|D_1)}\frac{P(M_\delta = 0)}{P(M_{i,\delta}=1)}\Big| M_\delta = 0\Big).
\end{aligned} \tag{18}$$

The result follows combining Equations (18) and (14) □

6.5 Theorem 5

Using the same steps as in the proof of Theorem 1, one can show that for all $t \geq 2$,

$$\delta_{1t} = E\left(\frac{Y_t - Y_{t-1} - E(Y_t - Y_{t-1}|D_{t-1}, M_t = 0)}{D_t - D_{t-1}}\Big| M_t = 1\right).$$

This proves the result □

6.6 Theorem 6

Using the same steps as in the proof of Theorem 1, one can show that for all $t \geq 2$,

$$\delta_{2it} = \frac{E\left(Y_t - Y_{t-1} | M_{i,t} = 1\right) - E\left(\left(Y_t - Y_{t-1}\right) \frac{P(M_{i,t}=1|D_{t-1})}{P(M_t=0|D_{t-1})} \frac{P(M_t=0)}{P(M_{i,t}=1)} \middle| M_t = 0\right)}{E(D_t - D_{t-1} | M_{i,t} = 1)},$$
$$\delta_{2dt} = \frac{E\left(Y_t - Y_{t-1} | M_{d,t} = 1\right) - E\left(\left(Y_t - Y_{t-1}\right) \frac{P(M_{d,t}=1|D_{t-1})}{P(M_t=0|D_{t-1})} \frac{P(M_t=0)}{P(M_{d,t}=1)} \middle| M_t = 0\right)}{E(D_{t-1} - D_t | M_{d,t} = 1)}.$$

This proves the result □

6.7 Theorem 7

Using the same reasoning as for Theorem 1, we have, for all $(d_{t-1}, d_t) \in \text{Supp}(D_{t-1}, D_t | M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0)$,

$$\begin{aligned}
&E\left[Y_{t+\ell} - Y_{t-1} | D_{t-1} = d_{t-1}, M_t = 0, \sum_{k=1}^{\ell} M_{t+k} = 0\right] \\
=&E\left[Y_{t+\ell}(d_{t-1}) - Y_{t-1}(d_{t-1}) | D_{t-1} = D_t = D_{t+\ell} = d_{t-1}\right] \\
=&E\left[Y_{t+\ell}(d_{t-1}) - Y_{t-1}(d_{t-1}) | D_{t-1} = d_{t-1}, D_t = D_{t+\ell} = d_t\right] \\
=&E\left[Y_{t+\ell}(d_{t-1}) - Y_{t-1}(d_{t-1}) | D_{t-1} = d_{t-1}, D_t = d_t, M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0\right],
\end{aligned}$$

where the second equality follows from Assumption 10. Then, reasoning again as in Theorem 1, we obtain

$$E\left(\frac{Y_{t+\ell} - Y_{t-1} - E\left(Y_{t+\ell} - Y_{t-1} | D_{t-1}, M_t = 0, \sum_{k=1}^{\ell} M_{t+k} = 0\right)}{D_t - D_{t-1}} \middle| M_t = 1, \sum_{k=1}^{\ell} M_{t+k} = 0\right) = \delta_{1t:t+\ell}.$$

The result follows.

6.8 Theorem 8

We start by proving Equation (7). For all $t \geq \ell + 2$, for every d_1 in the support of $D_1 | F = t - \ell$,

$$\begin{aligned}
E(Y_t - Y_{t-\ell-1} | D_1 = d_1, F > t) =& E(Y_t(\boldsymbol{d}_1) - Y_{t-\ell-1}(\boldsymbol{d}_1) | D_1 = ... = D_t = d_1) \\
=& E(Y_t(\boldsymbol{d}_1) - Y_{t-\ell-1}(\boldsymbol{d}_1)) \\
=& E(Y_t(\boldsymbol{d}_1) - Y_{t-\ell-1}(\boldsymbol{d}_1) | D_1 = d_1, F = t - \ell). \qquad (19)
\end{aligned}$$

It follows from Assumption 15 that d_1 belongs to the support of $D_1 | F = t - \ell$, so $E(Y_t - Y_{t-\ell-1} | D_1 = d_1, F > t)$ is well defined. The first equality follows from the definition of F.

The second and third equalities follow from Assumptions 13 and 14 and the law of iterated expectations.

Then,

$$\begin{aligned}
&E\left(Y_t - Y_{t-\ell-1} - E(Y_t - Y_{t-\ell-1}|D_1, F > t)|F = t-\ell\right)\\
=&E\left(Y_t - Y_{t-\ell-1} - E(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|D_1, F = t-\ell)|F = t-\ell\right)\\
=&E\left(Y_t(\boldsymbol{D}) - Y_{t-\ell-1}(\boldsymbol{D}_1) - E(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|D_1, F = t-\ell)|F = t-\ell\right)\\
=&E\left(E(Y_t(\boldsymbol{D}) - Y_{t-\ell-1}(\boldsymbol{D}_1) - Y_t(\boldsymbol{D}_1) + Y_{t-\ell-1}(\boldsymbol{D}_1)|D_1, F = t-\ell)|F = t-\ell\right)\\
=&E\left(Y_t(\boldsymbol{D}) - Y_t(\boldsymbol{D}_1)|F = t-\ell\right).
\end{aligned}$$

The first equality follows from (19). The second equality follows from the definition of F and Assumption 13. The third and fourth equalities follow from the law of iterated expectations. This proves the result □

We now prove Equation (8). It follows from the definition of F and Assumption 13 that

$$E\left(Y_t - Y_{t-\ell-1}|F = t-\ell\right) = E\left(Y_t(\boldsymbol{D}) - Y_{t-\ell-1}(\boldsymbol{D}_1)|F = t-\ell\right).$$

Then,

$$\begin{aligned}
&E\quad (Y_t - Y_{t-\ell-1})\frac{P(F = t-\ell|D_1)}{P(F > t|D_1)}\frac{P(F > t)}{P(F = t-\ell)}\bigg|F > t\bigg)\\
=&E\left(E(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|D_1, F = t-\ell)\frac{P(F = t-\ell|D_1)}{P(F > t|D_1)}\frac{P(F > t)}{P(F = t-\ell)}\bigg|F > t\right)\\
=&E\left(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|F = t-\ell\right).
\end{aligned}$$

The first equality follows from the law of iterated expectations and (19). The second equality follows from the same steps as in the proof of Theorem 2. Combining the two previous displays proves the result □

References

Abadie, A. (2005), 'Semiparametric difference-in-differences estimators', *Review of Economic Studies* **72**(1), 1–19.

Abbring, J. H. and Van den Berg, G. J. (2003), 'The nonparametric identification of treatment effects in duration models', *Econometrica* **71**(5), 1491–1517.

Athey, S. and Imbens, G. W. (2006), 'Identification and inference in nonlinear difference-in-differences models', *Econometrica* **74**(2), 431–497.

Borusyak, K., Jaravel, X. and Spiess, J. (2021), Revisiting event study designs: Robust and efficient estimation. Working Paper.

Botosaru, I. and Gutierrez, F. H. (2018), 'Difference-in-differences when the treatment status is observed in only one period', *Journal of Applied Econometrics* **33**(1), 73–90.

Callaway, B., Goodman-Bacon, A. and Sant'Anna, P. H. (2021), 'Difference-in-differences with a continuous treatment', *arXiv preprint arXiv:2107.02637* .

Callaway, B. and Sant'Anna, P. H. (2020), 'Difference-in-differences with multiple time periods', *Journal of Econometrics* .

de Chaisemartin, C. and D'Haultfœuille, X. (2018), 'Fuzzy differences-in-differences', *The Review of Economic Studies* **85**(2), 999–1028.

de Chaisemartin, C. and D'Haultfœuille, X. (2020*a*), 'Difference-in-differences estimators of intertemporal treatment effects', *Available at SSRN 3731856* .

de Chaisemartin, C. and D'Haultfœuille, X. (2020*b*), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–2996.

de Chaisemartin, C., D'Haultfœuille, X. and Guyonvarch, Y. (2019), 'Fuzzy differences-in-differences with stata', *Stata Journal* **19**(2), 435–458.

Deschênes, O. and Greenstone, M. (2007), 'The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather', *American economic review* **97**(1), 354–385.

D'Haultfoeuille, X., Hoderlein, S. and Sasaki, Y. (2021), Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. arXiv preprint arXiv:2104.14458.

Fajgelbaum, P. D., Goldberg, P. K., Kennedy, P. J. and Khandelwal, A. K. (2020), 'The return to protectionism', *The Quarterly Journal of Economics* **135**(1), 1–55.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* .

Graham, B. S. and Powell, J. L. (2012), 'Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models', *Econometrica* **80**(5), 2105–2152.

Hoderlein, S. and White, H. (2012), 'Nonparametric identification in nonseparable panel data models with generalized fixed effects', *Journal of Econometrics* **168**(2), 300–314.

Houngbedji, K. (2016), 'ABSDID: Stata module to estimate treatment effect with Abadie semi-parametric DID estimator'.
URL: *https://ideas.repec.org/c/boc/bocode/s458134.html*

Malani, A. and Reif, J. (2015), 'Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform', *Journal of Public Economics* **124**, 1–17.

Rios-Avila, F., Sant'Anna, P. and Callaway, B. (2021), 'Csdid: Stata module for the estimation of difference-in-difference models with multiple time periods'.
URL: *https://EconPapers.repec.org/RePEc:boc:bocode:s458976*

Robins, J. (1986), 'A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect', *Mathematical modelling* **7**(9-12), 1393–1512.

Sun, L. and Abraham, S. (2020), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **forthcoming**.